Analyzing and Communicating Scientific Data with Visual Analytics

Enrico Bertini, *Assistant Professor*

NYU | POLYTECHNIC SCHOOL OF ENGINEERING

(NASA JPL - June 21, 2015)

I guess I don’t need to tell NASA we live in the “Data Era” right? :)

We all understand data has lots of value ...

… but how do we extract useful information and knowledge out of data?

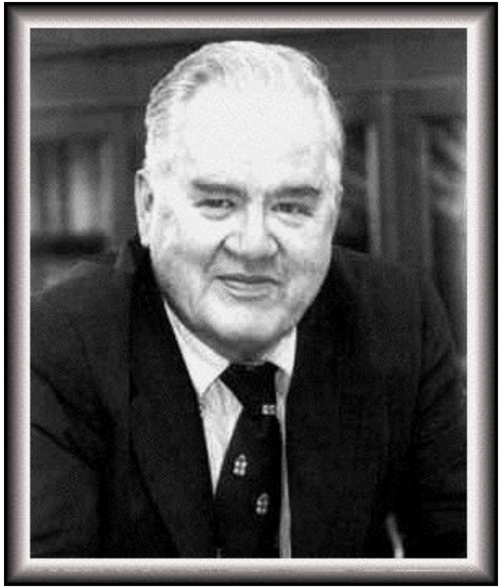
… and how do we communicate this information *effectively*, *truthfully*, and *persuasively* to others?

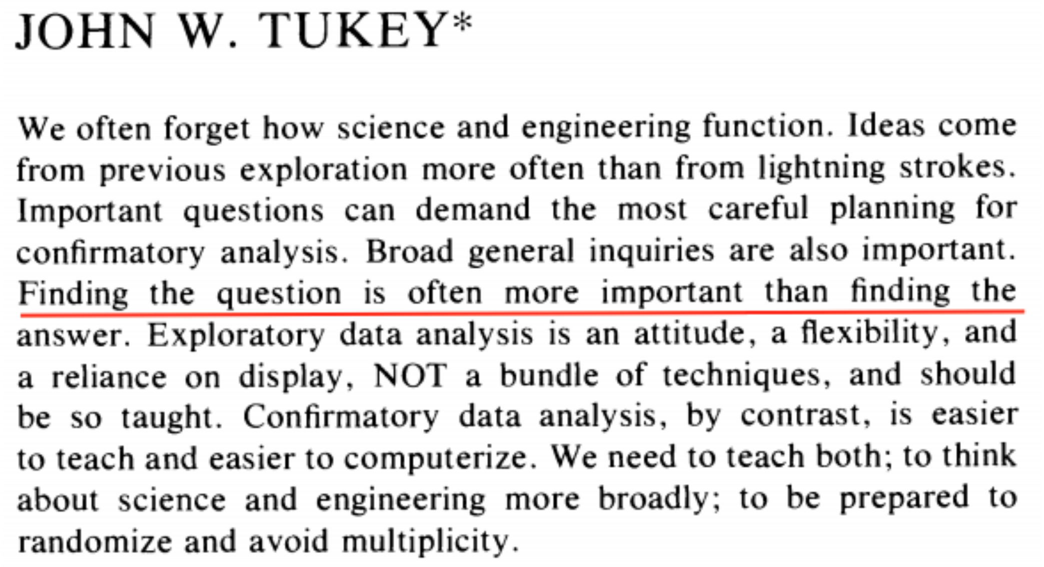
Traditional scientific process:

1) Formulate a question first.
2) Collect necessary data.
3) Run experiment to answer the question.

… when data largely available/easy to produce:

1) We got data! What shall we do with it?
2) Let’s look into it.
3) Mmm … How?

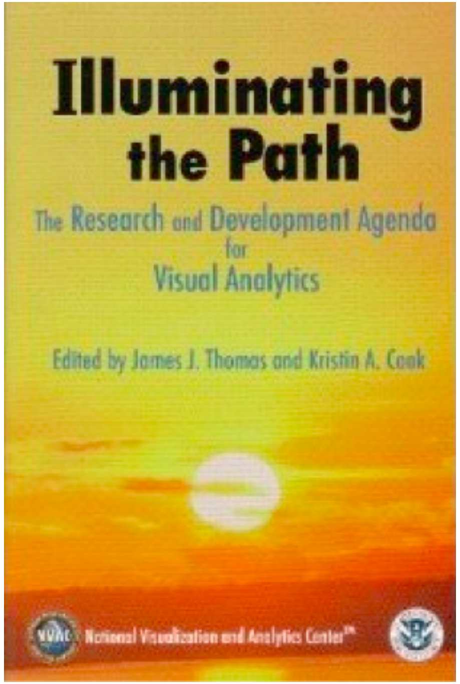


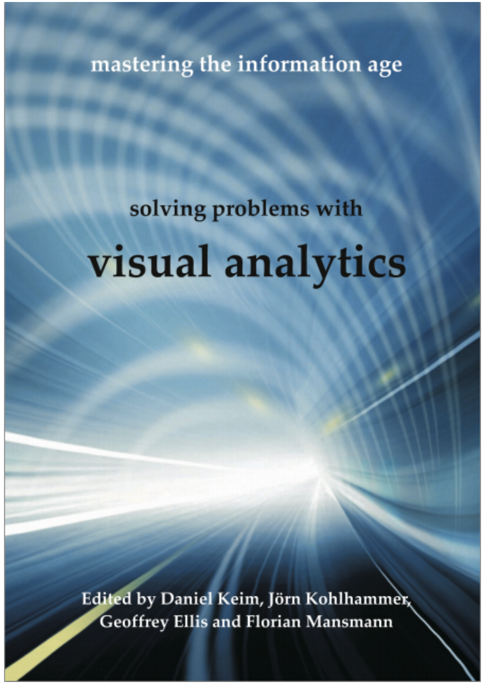


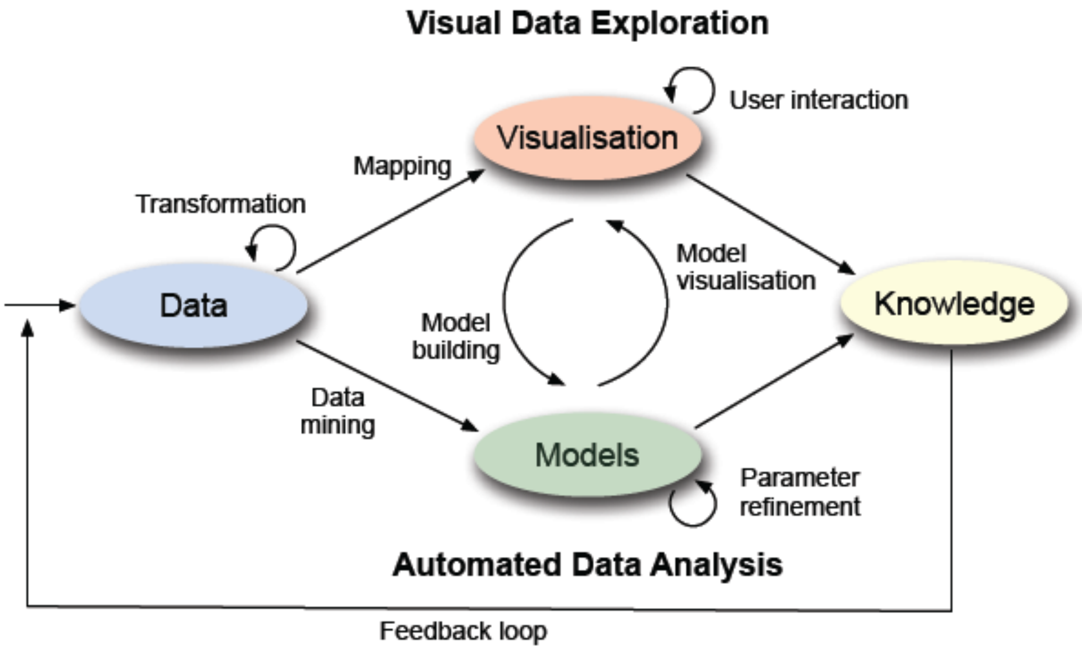
JOHN W. TUKEY*

We often forget how science and engineering function. Ideas come from previous exploration more often than from lightning strokes. Important questions can demand the most careful planning for confirmatory analysis. Broad general inquiries are also important. Finding the question is often more important than finding the answer. Exploratory data analysis is an attitude, a flexibility, and a reliance on display, NOT a bundle of techniques, and should be so taught. Confirmatory data analysis, by contrast, is easier to teach and easier to computerize. We need to teach both; to think about science and engineering more broadly; to be prepared to randomize and avoid multiplicity.

Visual Analytics: “The science of analytical reasoning facilitated by interactive visual interfaces”


Illuminating the Path
The Research and Development Agenda for Visual Analytics
Edited by James J. Thomas and Kristin A. Cook
NVAC National Visualization and Analytics Center™


mastering the information age
solving problems with
visual analytics
Edited by Daniel Keim, Jörn Kohlhammer, Geoffrey Ellis and Florian Mansmann


Visual Data Exploration
User interaction
Visualisation
Mapping
Transformation
Model visualisation
Data
Knowledge
Model building
Data mining
Models
Parameter refinement
Automated Data Analysis
Feedback loop

Examples from our lab ...

a. Visual Analytics for Drug Discovery.

Assay Plates

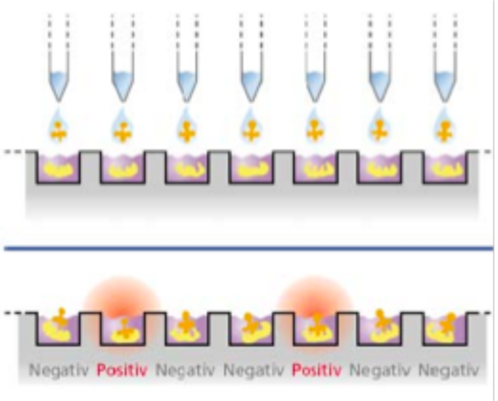
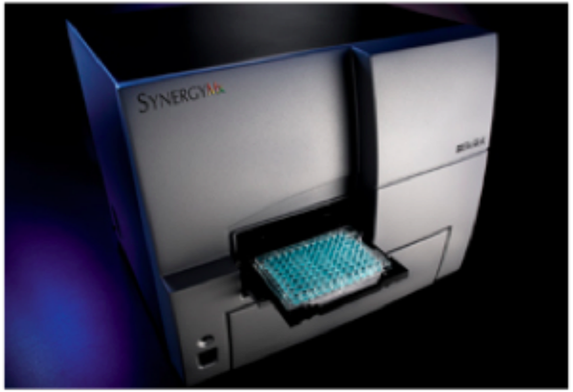
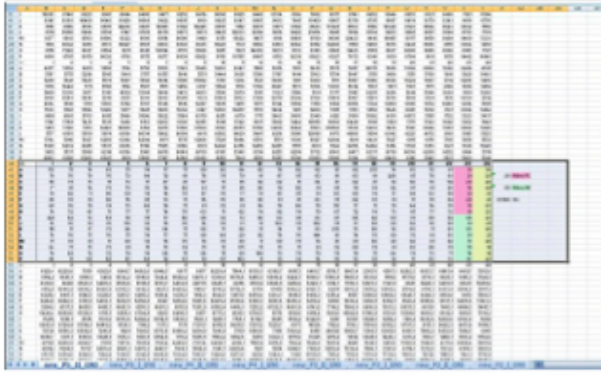


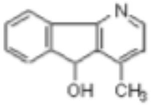
Plate Reader



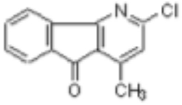
Data



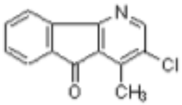
Structure-Activity Relationship (SAR) Analysis



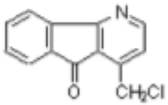
(VII)



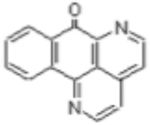
(VIII)



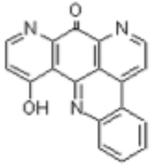
(IX)



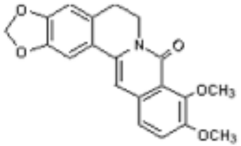
(X)



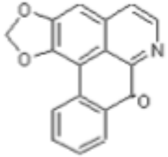
(XI)



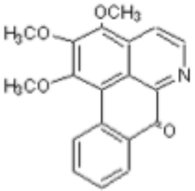
(XII)



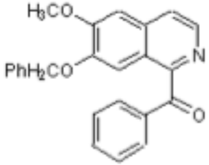
(XIII)



(XIV)

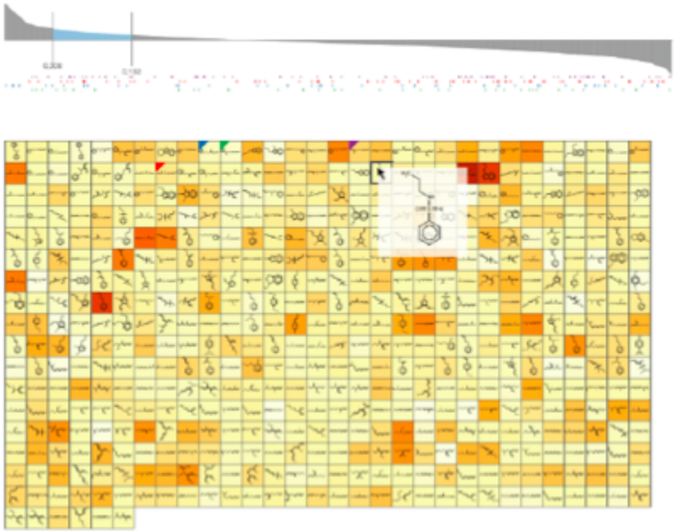


(XV)



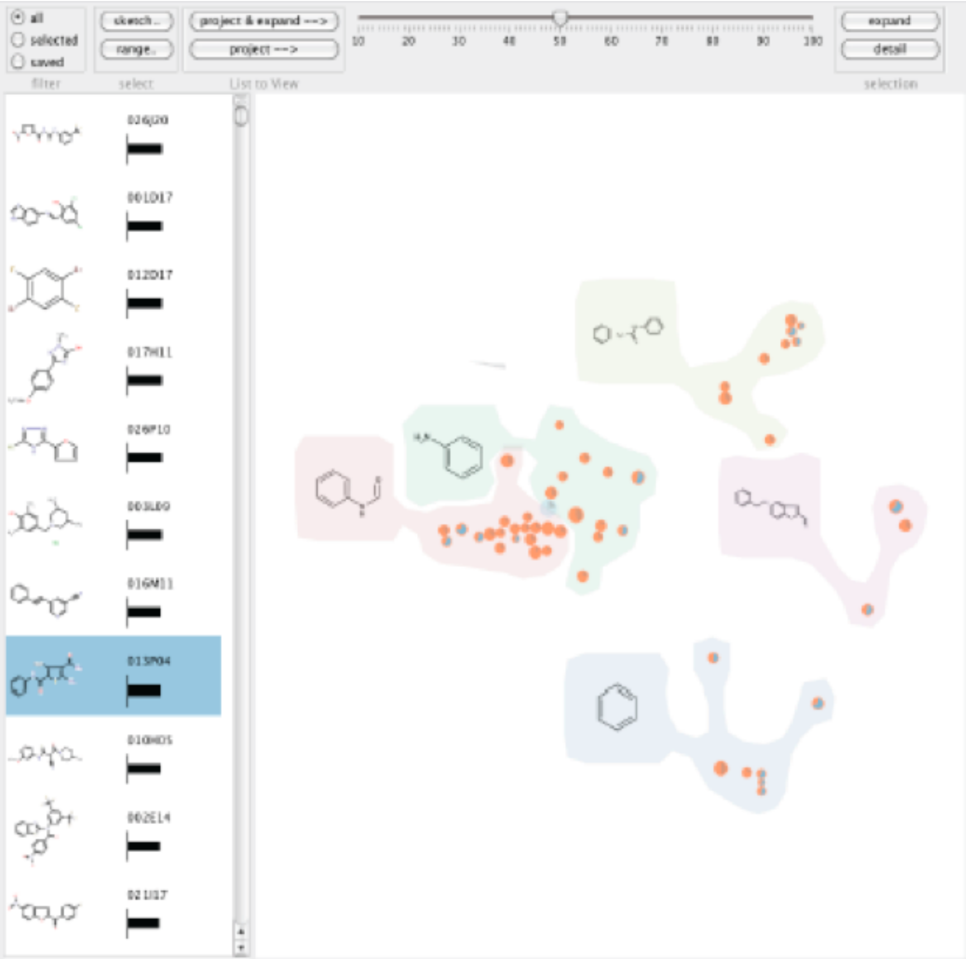
(XVI)

	O-S-C ‖ O	S-C-N ‖ O	S-C-N | C	S-Ċ-N'	C-N-C
O-S-C-N ‖ O	1	1	0	1	0
S-C-N-C ‖ O	0	1	1	1	1



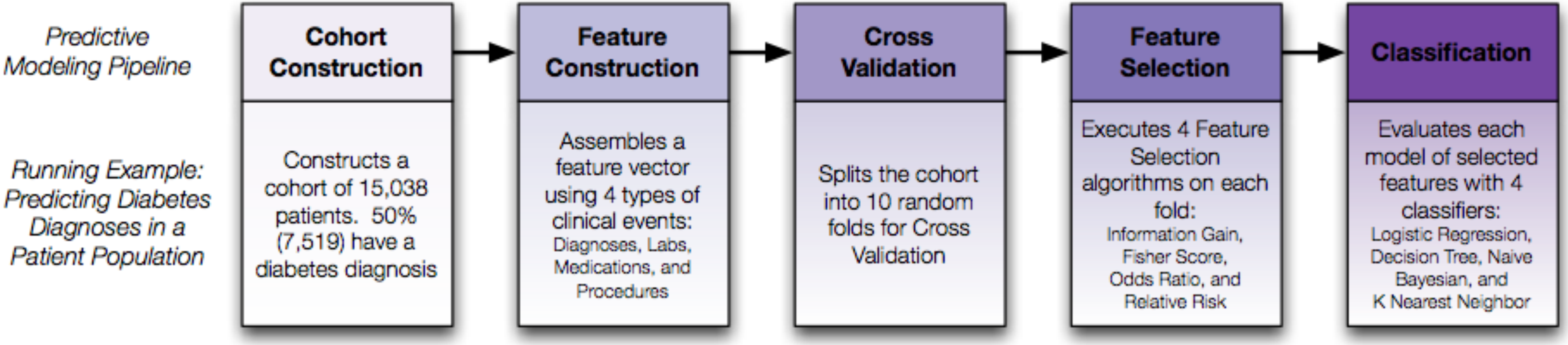
Mining algorithms necessary to extract meaningful molecular fragments.

HiTSEE



Bertini, Enrico, et al. "**HiTSEE: a visualization tool for hit selection and analysis in high-throughput screening experiments.**" Biological Data Visualization (BioVis), 2011 IEEE Symposium on. IEEE, 2011.

b. Visual Comparison of Machine Learning Models for Healthcare Analytics.



Predictive Modeling Pipeline

Running Example: Predicting Diabetes Diagnoses in a Patient Population

Cohort Construction

Constructs a cohort of 15,038 patients. 50% (7,519) have a diabetes diagnosis

Feature Construction

Assembles a feature vector using 4 types of clinical events: Diagnoses, Labs, Medications, and Procedures

Cross Validation

Splits the cohort into 10 random folds for Cross Validation

Feature Selection

Executes 4 Feature Selection algorithms on each fold: Information Gain, Fisher Score, Odds Ratio, and Relative Risk

Classification

Evaluates each model of selected features with 4 classifiers: Logistic Regression, Decision Tree, Naive Bayesian, and K Nearest Neighbor

(Work in collaboration with Adam Perer @ IBM Watson)

Parallel computation of multiple models

Feature Selection
(Information Gain, Fisher Score, Odds Ratio, Relative Risk, …)

X

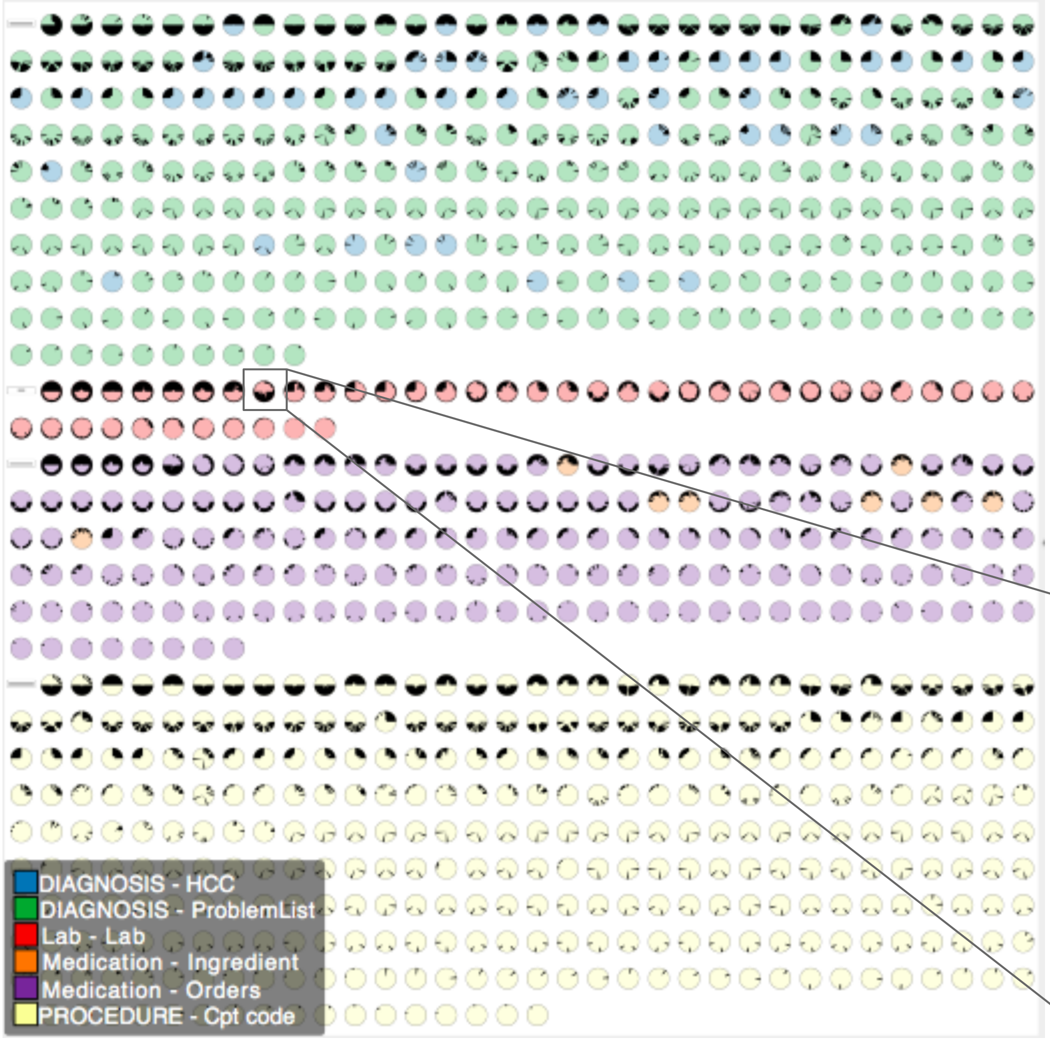
Classification
(Logistic Regression, Decision Trees, Naive Bayes, kNN, …)

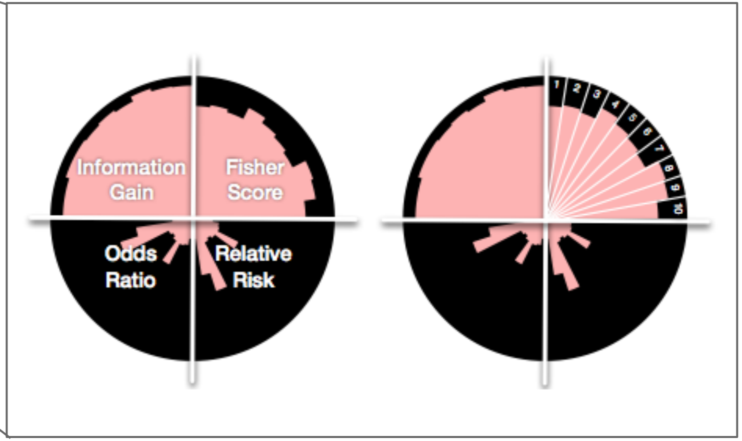
X

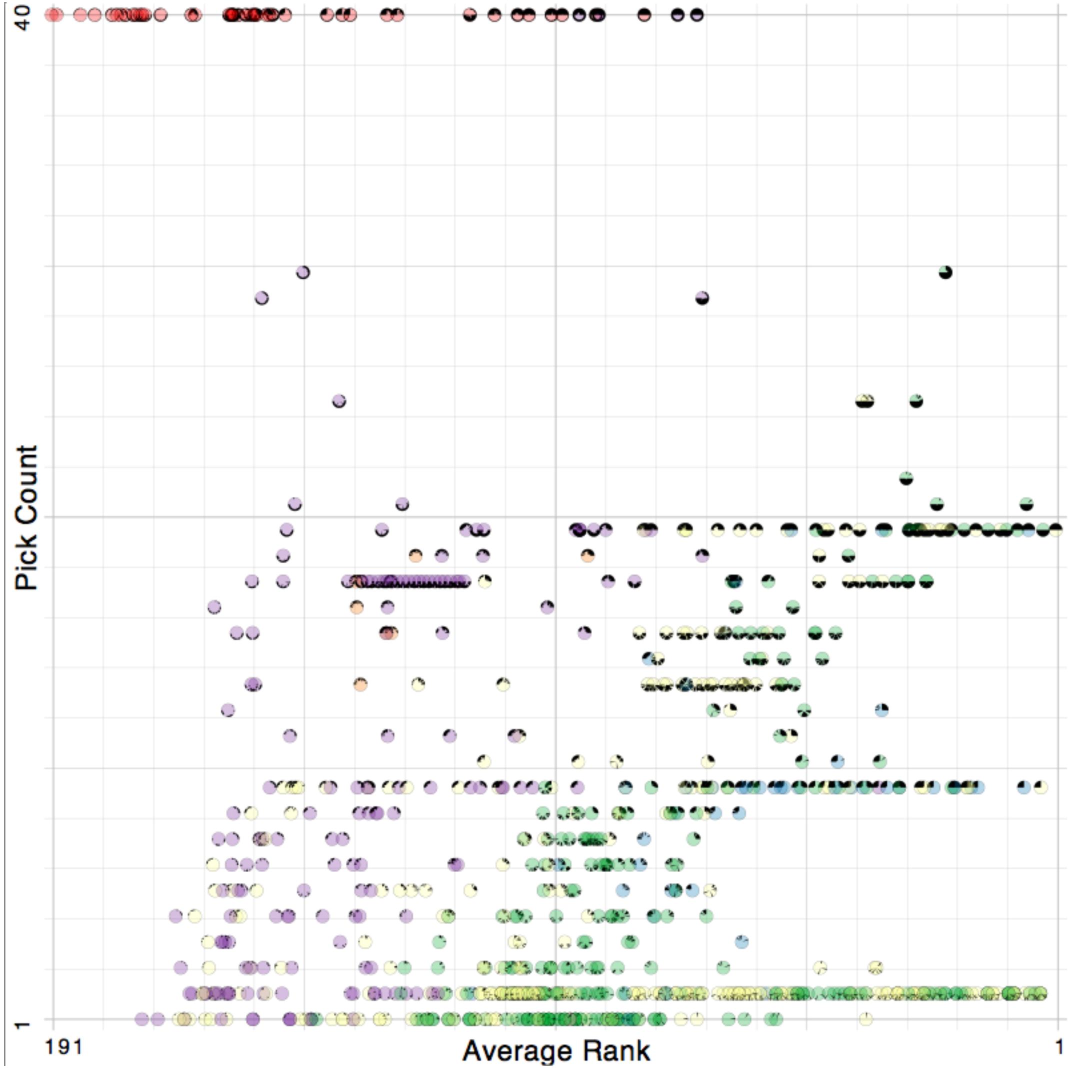
Folds (Samples)
10-folds validation

INFUSE

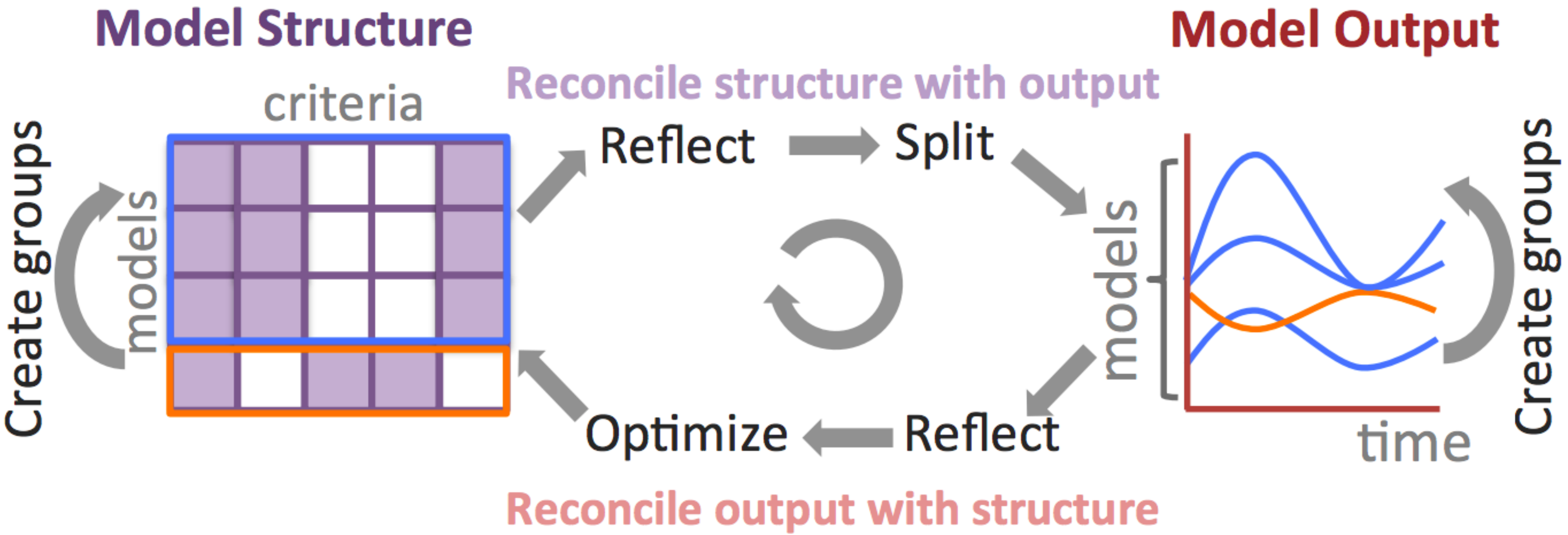
- Each dot is a feature (e.g., lab test)
- Each quadrant represents a feature selection algorithm
- Each segment represents a fold (sample)
- Length of the bar represents the ranking






40
Pick Count
1
191
Average Rank
1

c. Visual Reconciliation of Alternate Similarity Spaces in Climate Modeling.



Model Structure

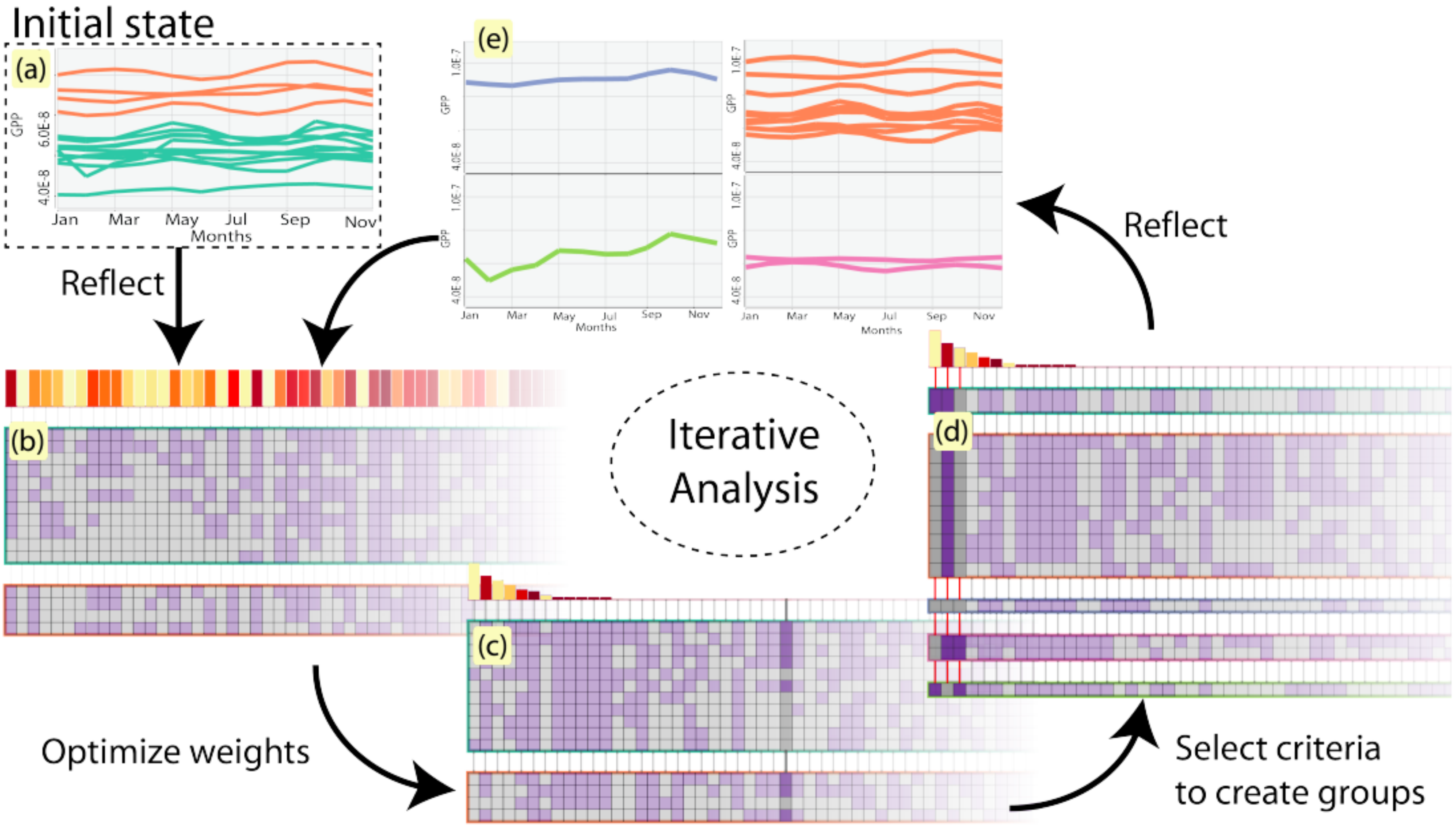
Model Output

Reconcile structure with output

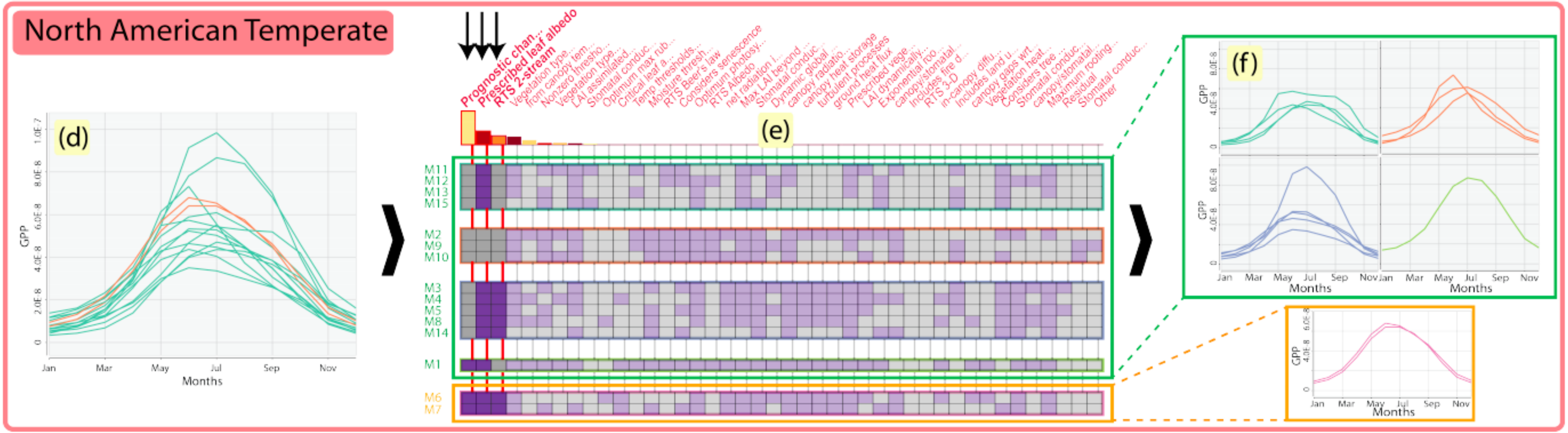
Reflect → Split

Optimize ← Reflect

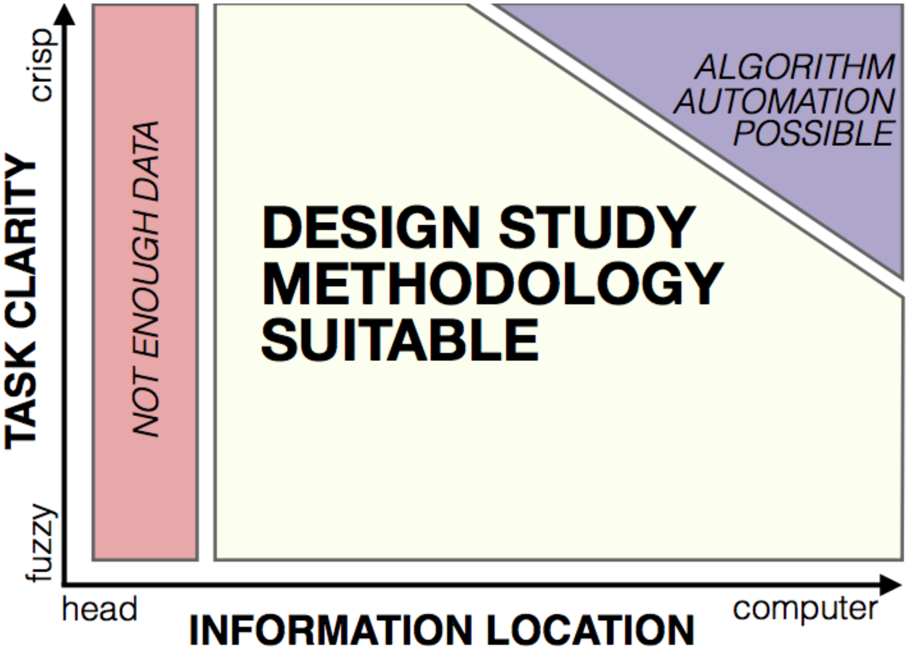
Reconcile output with structure



Initial state
(a)
GPP
6.0E-8
4.0E-8
Jan
Mar
May
Jul
Sep
Nov
Months
Reflect
(e)
1.0E-7
GPP
4.0E-8
Jan
Mar
May
Jul
Sep
Nov
Months
Reflect
(b)
Iterative Analysis
(d)
(c)
Optimize weights
Select criteria to create groups


North American Temperate
(d)
GPP
1.0E-7
8.0E-8
6.0E-8
4.0E-8
2.0E-8
0
Jan
Mar
May
Jul
Sep
Nov
Months
(e)
Prognostic chan...
Prescribed leaf albedo
RTS 2-stream
Vegetation type...
from canopy tem...
Nonzero thresho...
Vegetation type...
LAI assimilated...
Stomatal conduc...
Optimum max rub...
Critical leaf a...
Temp thresholds...
Moisture thresh...
RTS Beer's law
Considers senescence
Optimum photosy...
RTS Albedo
net radiation i...
Max LAI beyond ...
Stomatal conduc...
Dynamic global ...
canopy radiatio...
canopy heat storage
turbulent processes
ground heat flux
Prescribed vege...
LAI dynamically...
Exponential roo...
canopy/stomatal...
Includes fire d...
RTS 3-D
in-canopy diffu...
Includes land u...
canopy gaps wrt...
Vegetation heat...
Considers tree ...
Stomatal conduc...
canopy/stomatal...
Maximum rooting...
Residual
Stomatal conduc...
Other
M11
M12
M13
M15
M2
M9
M10
M3
M4
M5
M8
M14
M1
M6
M7
(f)
GPP
8.0E-8
4.0E-8
0
GPP
8.0E-8
4.0E-8
0
Jan
Mar
May
Jul
Sep
Nov
Months
Jan
Mar
May
Jul
Sep
Nov
Months
GPP
6.0E-8
4.0E-8
0
Jan
Mar
May
Jul
Sep
Nov
Months

Why use visualization?



Design Study Methodology: Reflections from the Trenches and the Stacks
Michael Sedlmair, Miriah Meyer, and Tamara Munzner
IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis), 18(12): 2431-2440, 2012.

Visualization can make complex problems trivial.

Let’s Play a Game! The Game of “15”

RULES

1) There are 2 players

2) Each player takes a digit in turn

3) Once a digit is taken, it cannot be used by any of the players again

4) The first player to get three digits that sum to 15 wins

{1, 2, 3, 4, 5, 6, 7, 8, 9}

Example taken from Prof. Pat Hanrahan’s EuroVis’09 keynote talk.

Tic-Tac-Toe: Herbert Simon’s “Problem Isomorph”

4	9	2
3	5	7
8	1	6

○ ✕

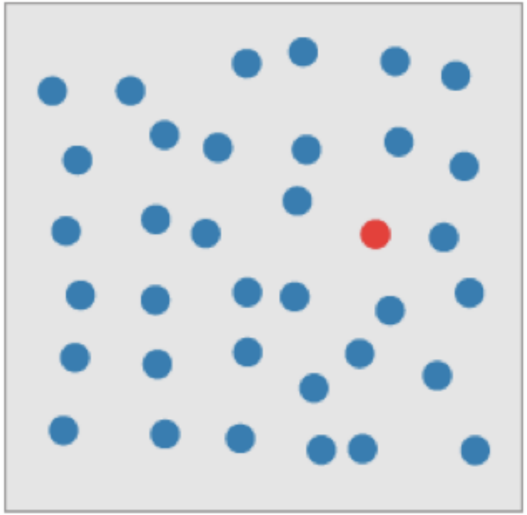
○ ✕

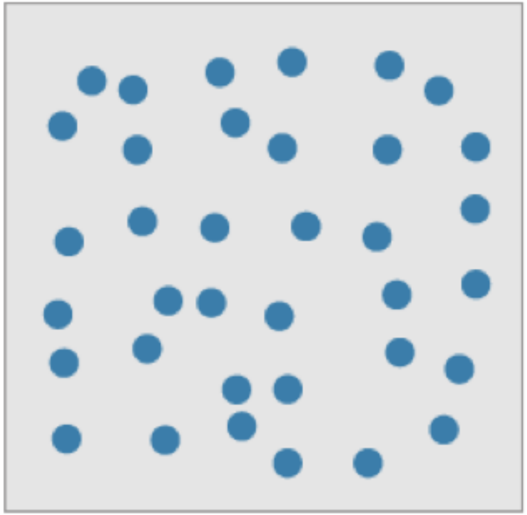
○ ✕

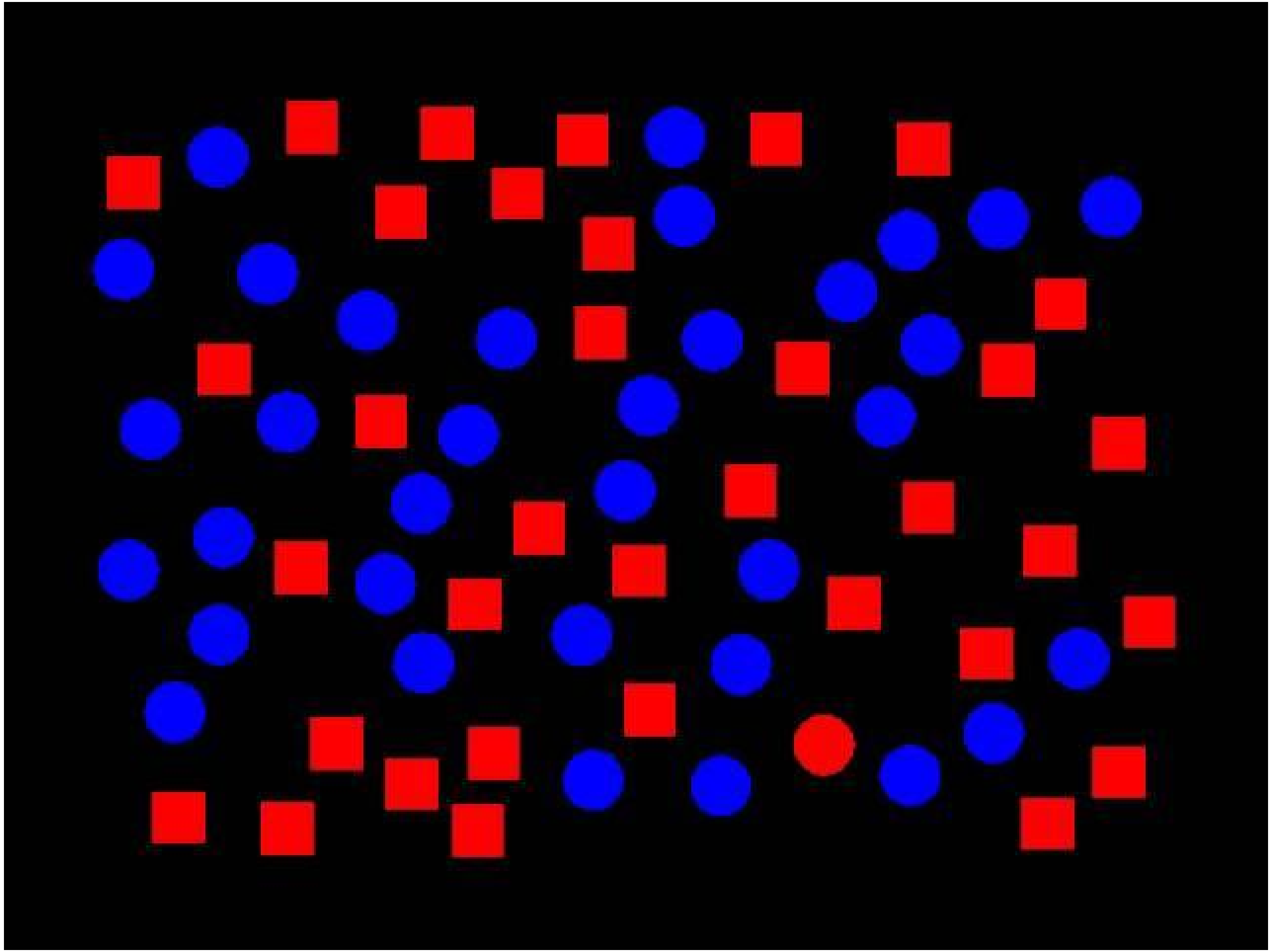
Visualization can be faster than your eyes can move!

Preattentive Processing

Preattentive features can be detected faster than eye movement (200 msec).



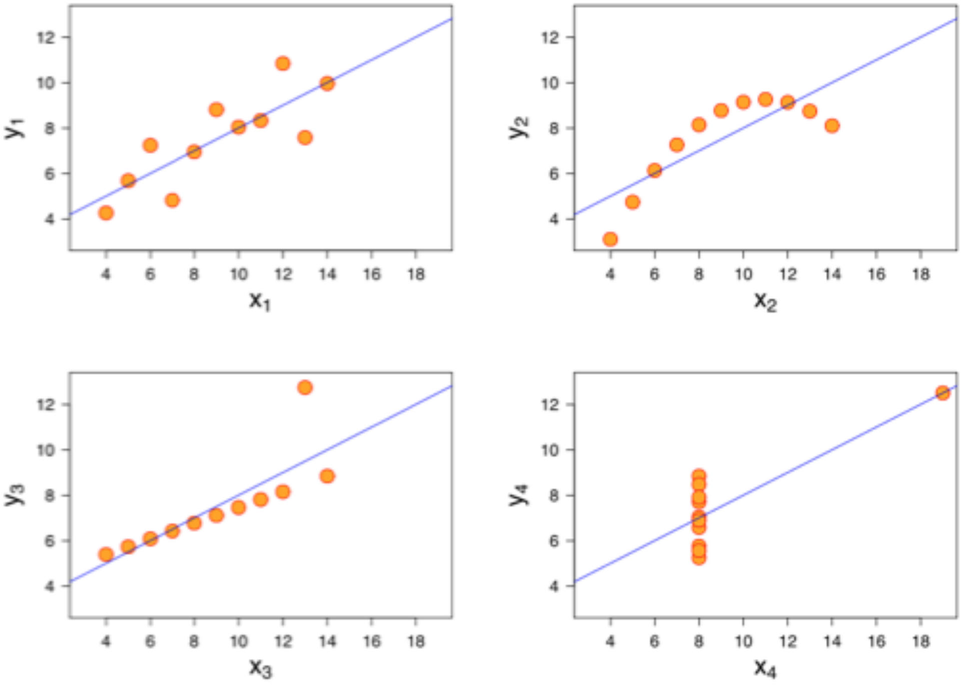


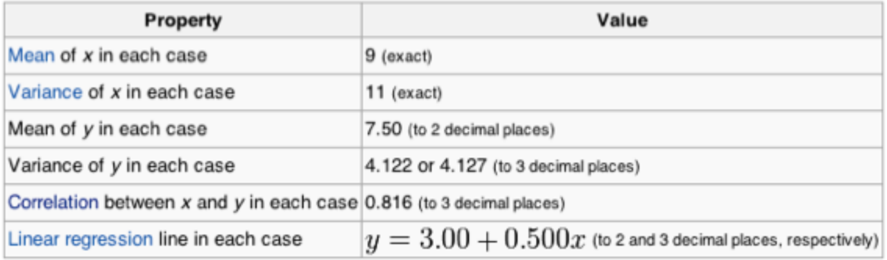


Visualization can reveal information that summary statistics may hide.

Anscombe’s Quartet

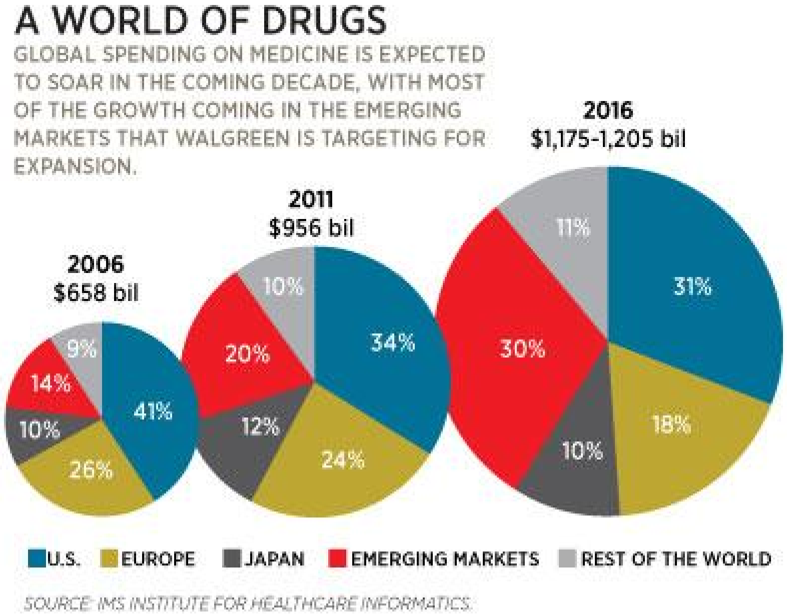
The risk of relying exclusively on numbers and statistics.





Property	Value
Mean of *x* in each case	9 (exact)
Variance of *x* in each case	11 (exact)
Mean of *y* in each case	7.50 (to 2 decimal places)
Variance of *y* in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between *x* and *y* in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)

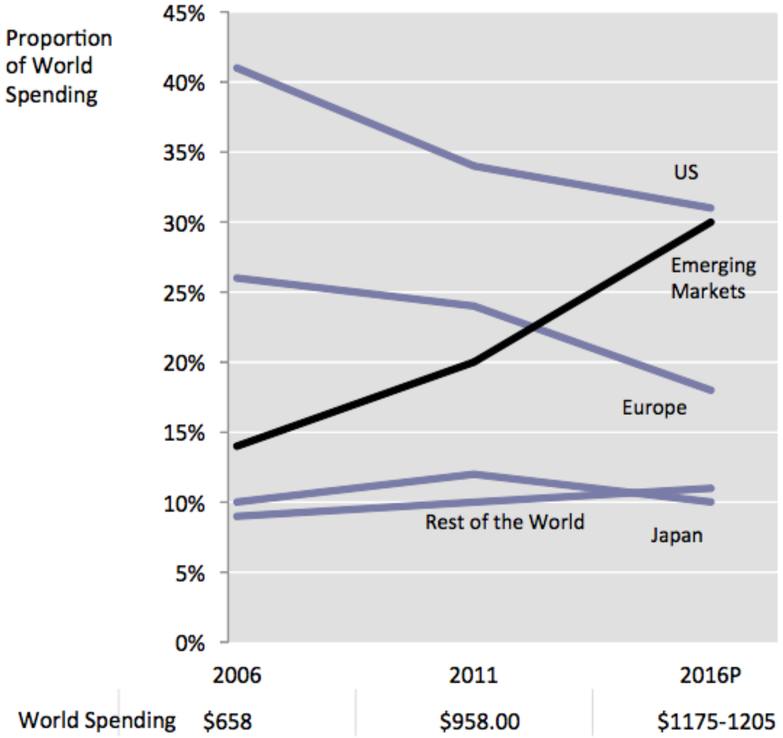
But … only if used properly!



A WORLD OF DRUGS

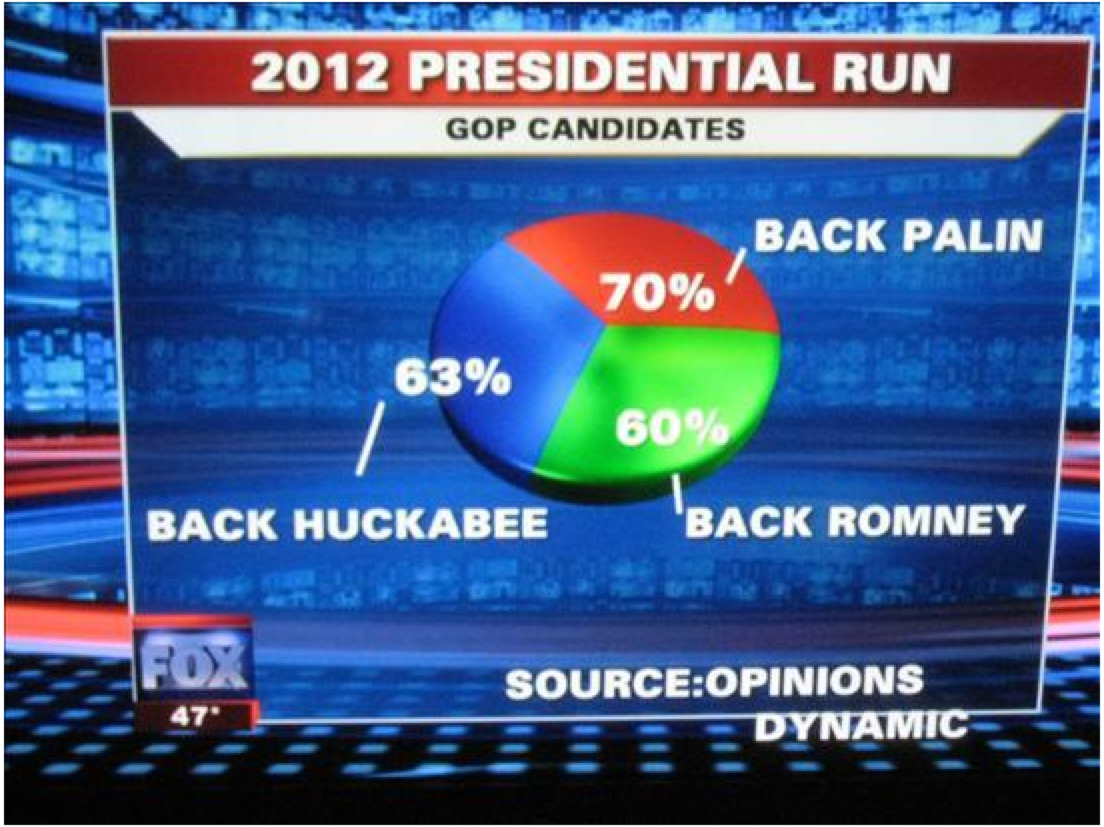
GLOBAL SPENDING ON MEDICINE IS EXPECTED TO SOAR IN THE COMING DECADE, WITH MOST OF THE GROWTH COMING IN THE EMERGING MARKETS THAT WALGREEN IS TARGETING FOR EXPANSION.

SOURCE: IMS INSTITUTE FOR HEALTHCARE INFORMATICS.

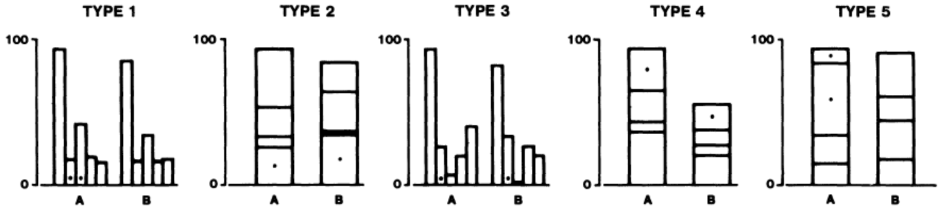


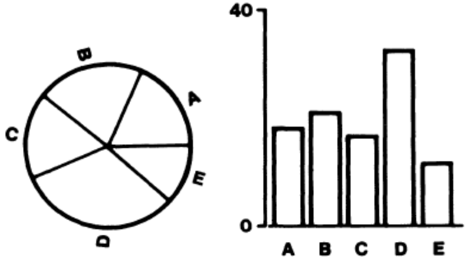
Example taken from: Junk Charts: Expanding circles of error

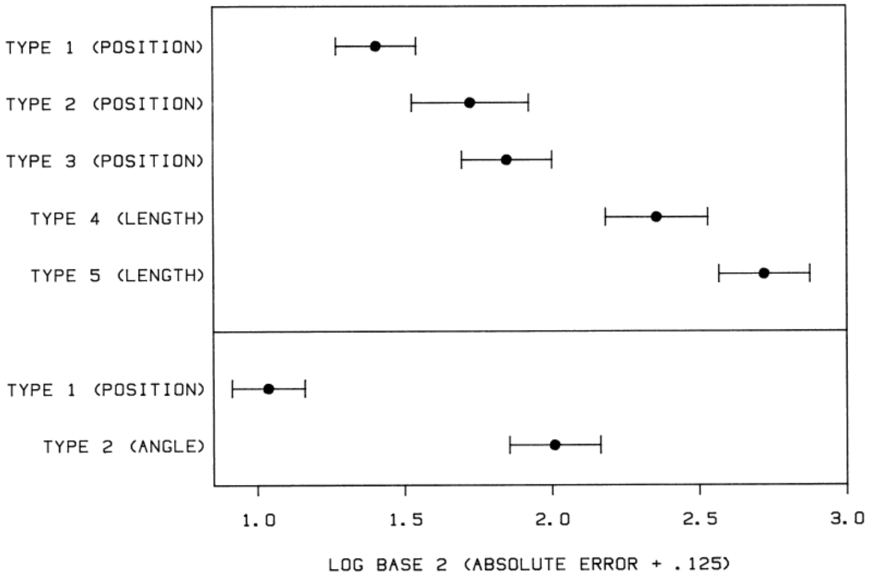
Some are plain wrong!



Graphical Perception

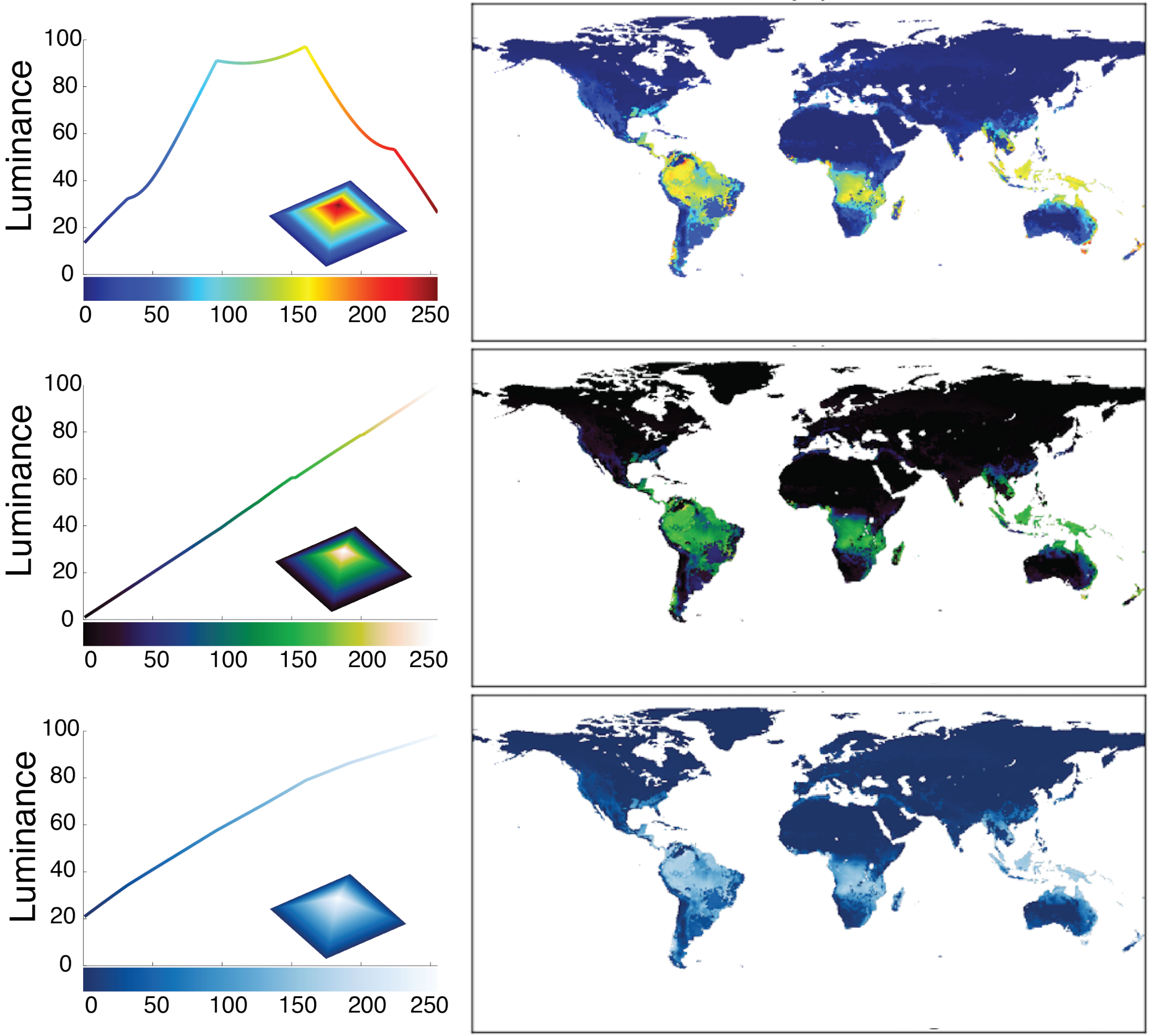


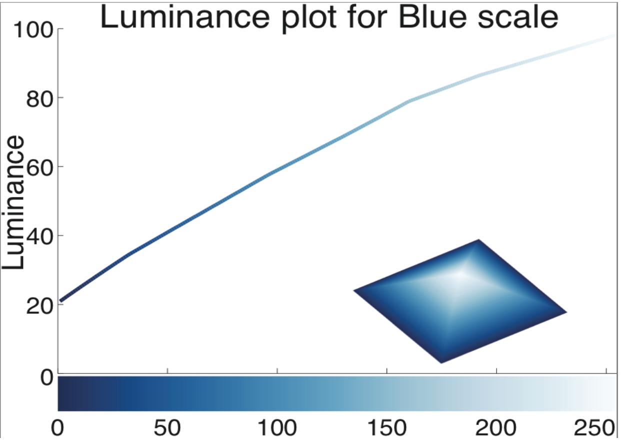




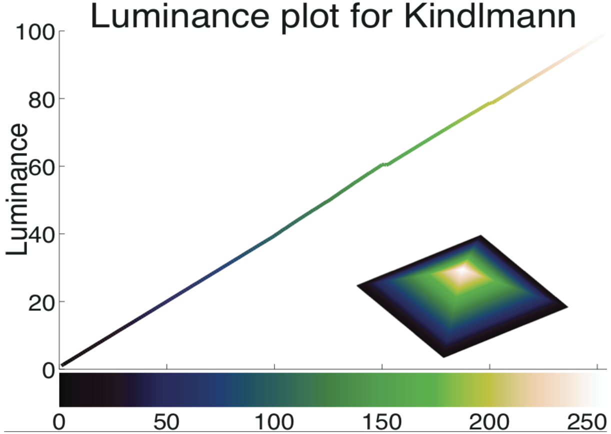
Cleveland, William S., and Robert McGill. "Graphical perception: Theory, experimentation, and application to the development of graphical methods."*Journal of the American Statistical Association* 79.387 (1984): 531-554.

Evaluation of Color Maps in Climate Science.

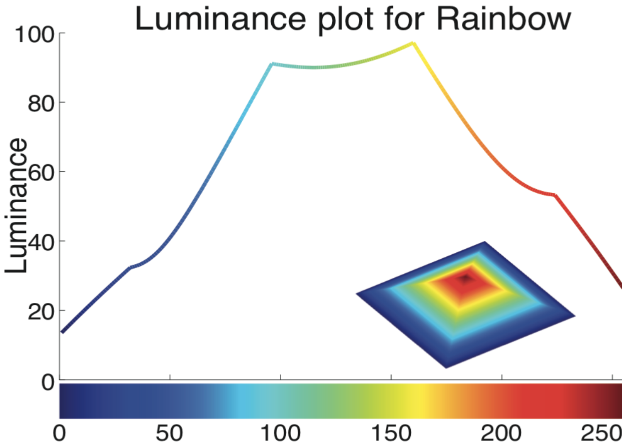

Luminance
100
80
60
40
20
0
0
50
100
150
200
250
Luminance
100
80
60
40
20
0
0
50
100
150
200
250
Luminance
100
80
60
40
20
0
0
50
100
150
200
250



Luminance plot for Blue scale



Luminance plot for Kindlmann



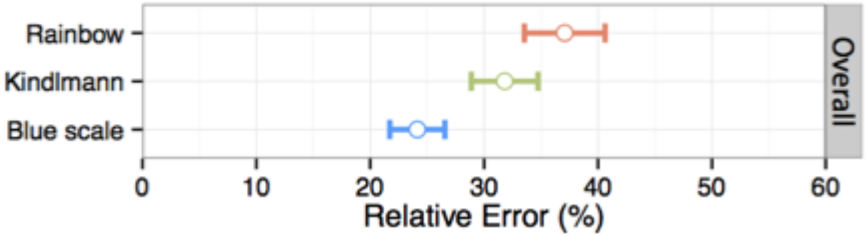
Luminance plot for Rainbow

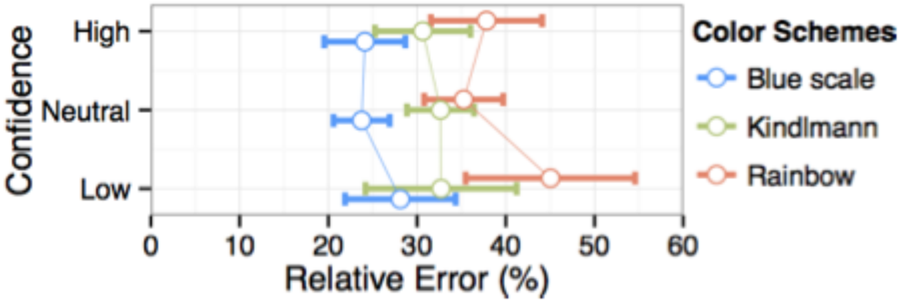
Task 1 - Magnitude Estimation

Task 2 - Similarity Estimation

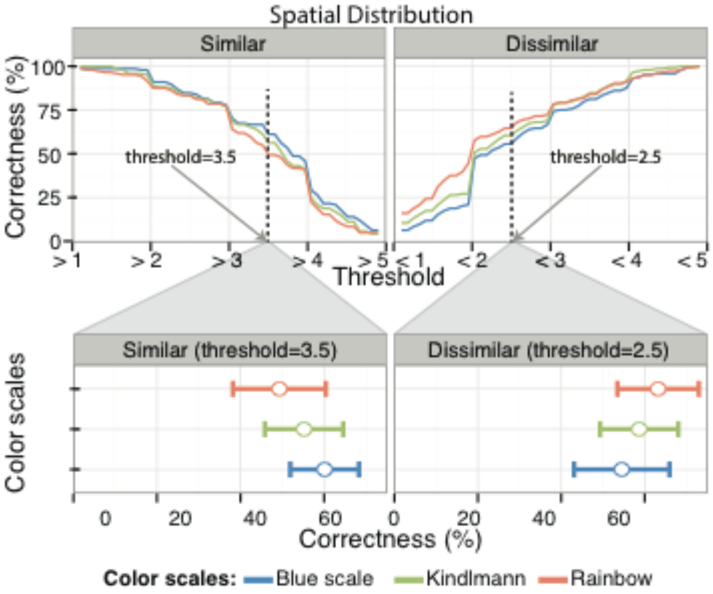
Task 3 - Area Identification

MAGNITUDE ESTIMATION

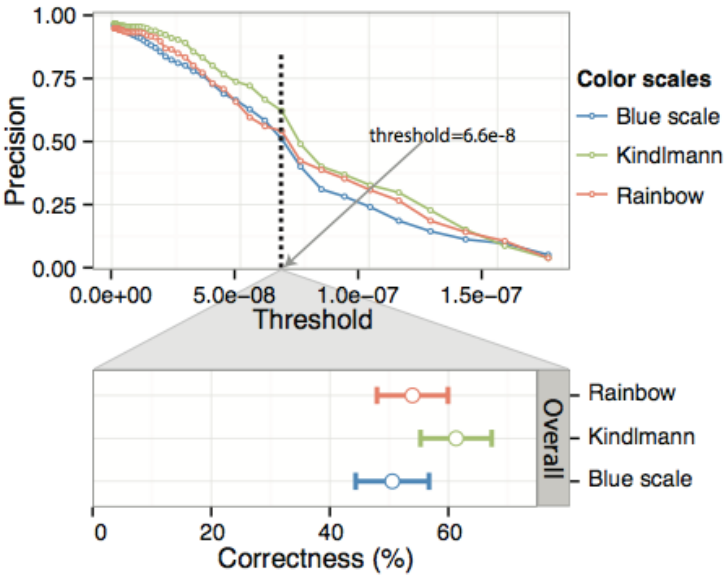




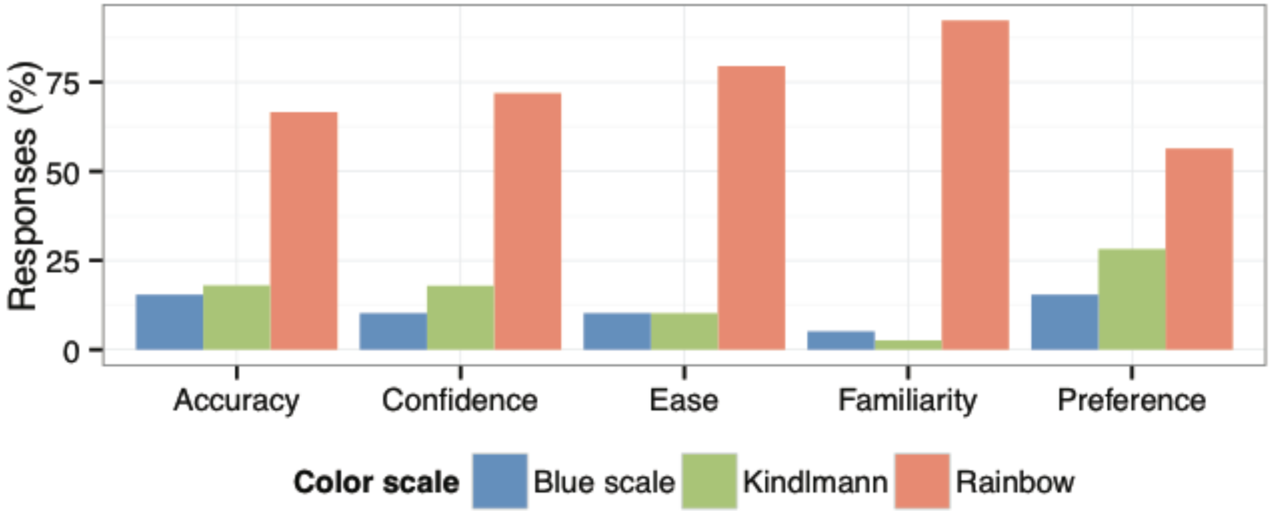
SIMILARITY ESTIMATION



AREA IDENTIFICATION



SUBJECTIVE PERCEPTION OF PERFORMANCE



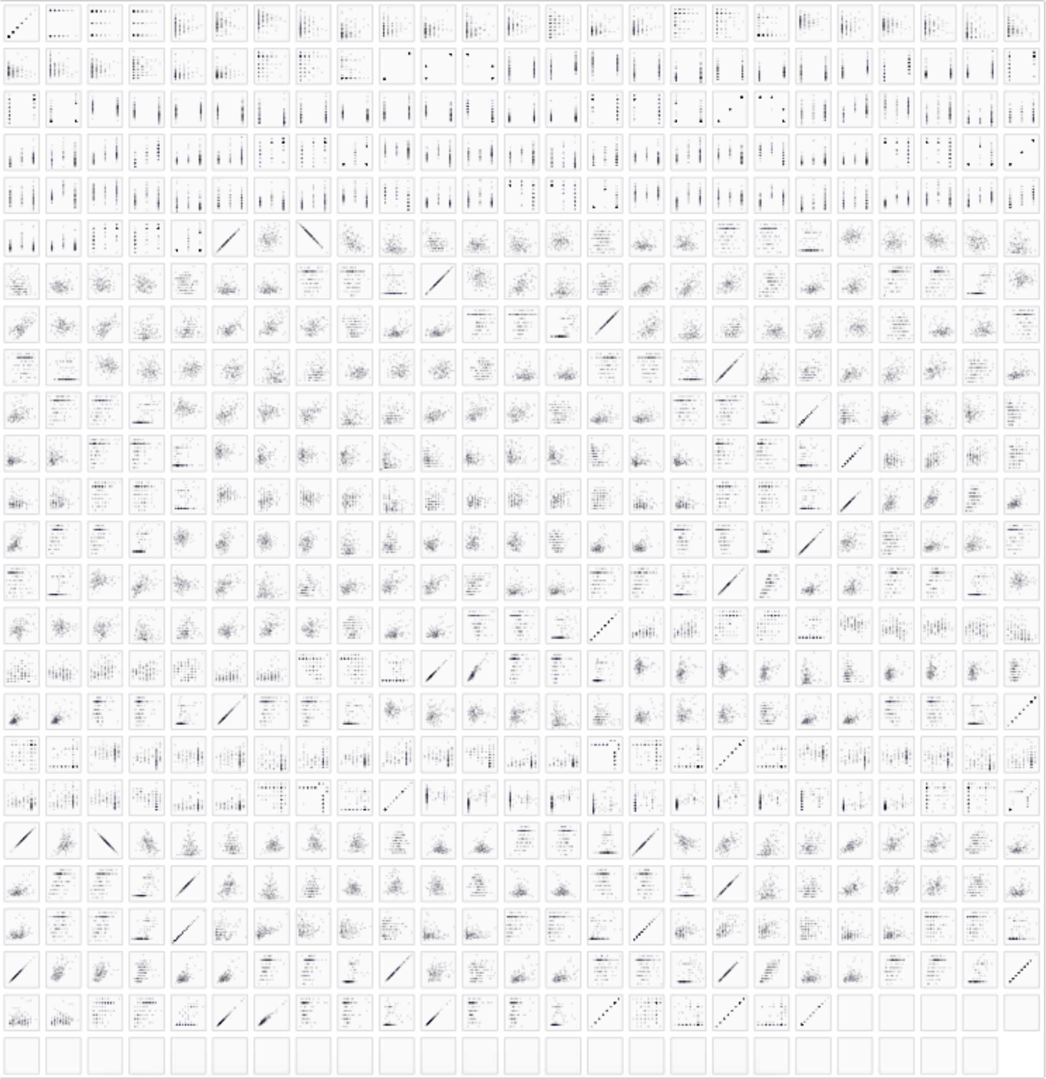
Selected Challenges

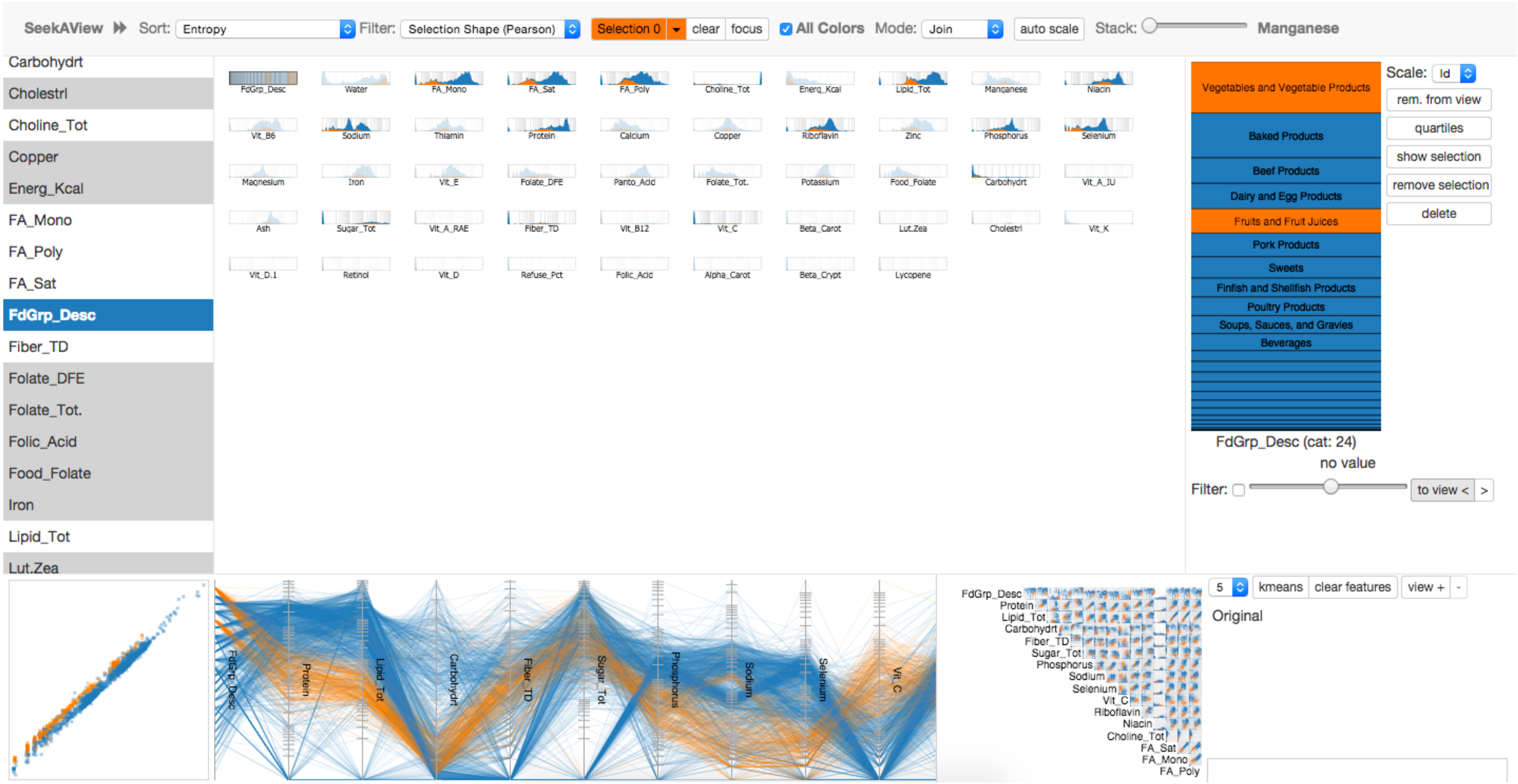
Automated and Interactive Methods.

High-Dimensional Data Spaces.

Evidence-Based Guidelines for Vis Design.

Sifting through a million plots.





SeekAView

Sort: Entropy

Filter: Selection Shape (Pearson)

Selection 0

☑ All Colors

Mode: Join

Stack:

Manganese

Scale: Id

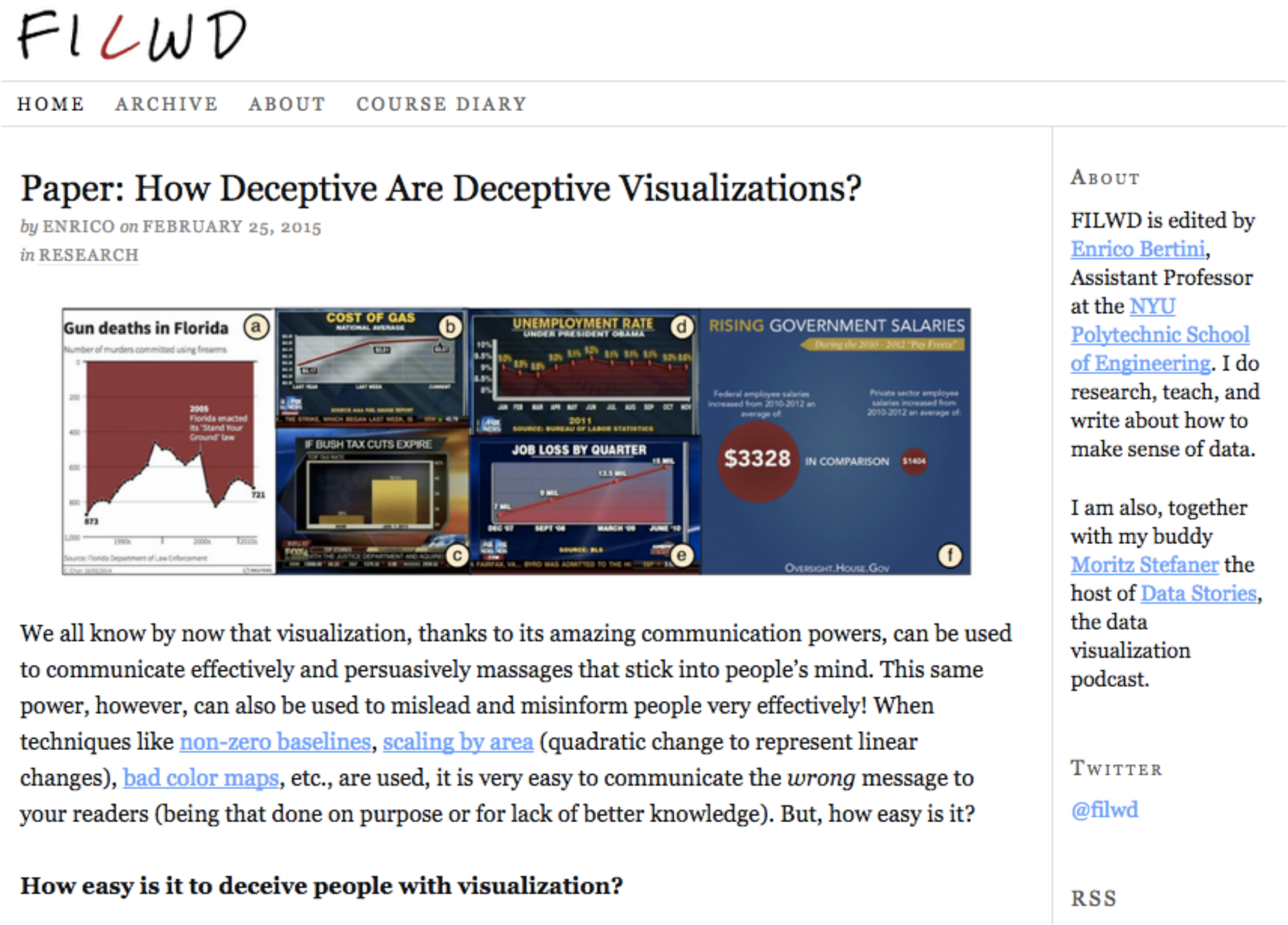
FdGrp_Desc (cat: 24)

no value

Filter: □

5

Original



FILWD

HOME ARCHIVE ABOUT COURSE DIARY

Paper: How Deceptive Are Deceptive Visualizations?

by ENRICO *on* FEBRUARY 25, 2015
in RESEARCH

We all know by now that visualization, thanks to its amazing communication powers, can be used to communicate effectively and persuasively massages that stick into people's mind. This same power, however, can also be used to mislead and misinform people very effectively! When techniques like non-zero baselines, scaling by area (quadratic change to represent linear changes), bad color maps, etc., are used, it is very easy to communicate the *wrong* message to your readers (being that done on purpose or for lack of better knowledge). But, how easy is it?

How easy is it to deceive people with visualization?

ABOUT

FILWD is edited by Enrico Bertini, Assistant Professor at the NYU Polytechnic School of Engineering. I do research, teach, and write about how to make sense of data.

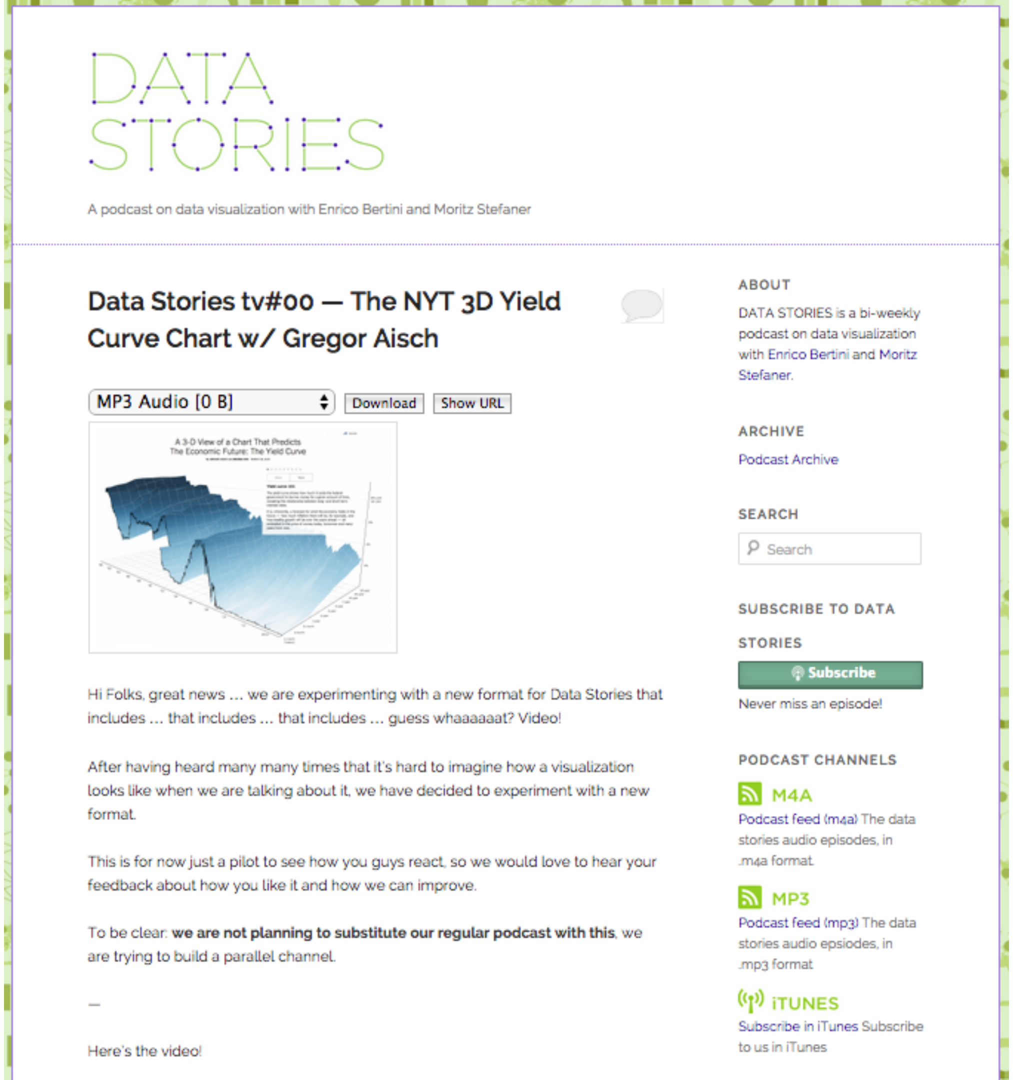
I am also, together with my buddy Moritz Stefaner the host of Data Stories, the data visualization podcast.

TWITTER

@filwd

RSS

http://fellinlovewithdata.com



A 3-D View of a Chart That Predicts
The Economic Future: The Yield Curve

DATA STORIES

A podcast on data visualization with Enrico Bertini and Moritz Stefaner

Data Stories tv#00 — The NYT 3D Yield Curve Chart w/ Gregor Aisch

MP3 Audio [0 B] | Download | Show URL

Hi Folks, great news … we are experimenting with a new format for Data Stories that includes … that includes … that includes … guess whaaaaaat? Video!

After having heard many many times that it's hard to imagine how a visualization looks like when we are talking about it, we have decided to experiment with a new format.

This is for now just a pilot to see how you guys react, so we would love to hear your feedback about how you like it and how we can improve.

To be clear: **we are not planning to substitute our regular podcast with this**, we are trying to build a parallel channel.

—

Here's the video!

ABOUT

DATA STORIES is a bi-weekly podcast on data visualization with Enrico Bertini and Moritz Stefaner.

ARCHIVE

Podcast Archive

SEARCH

Search

SUBSCRIBE TO DATA STORIES

Subscribe

Never miss an episode!

PODCAST CHANNELS

M4A

Podcast feed (m4a) The data stories audio episodes, in .m4a format

MP3

Podcast feed (mp3) The data stories audio epsiodes, in .mp3 format

iTUNES

Subscribe in iTunes Subscribe to us in iTunes

Thanks! Questions?

enrico.bertini@nyu.edu

http://enrico.bertini.io

@filwd